Applying Knowledge Discovery in Databases in Public Health Data Set: Challenges and Concerns

Kanittha Volrathongchia M.S. University of Wisconsin-Madison School of Nursing

Abstract. In attempting to apply Knowledge Discovery in Databases (KDD) to generate a predictive model from a health care dataset that is currently available to the public, the first step is to pre-process the data to overcome the challenges of missing data, redundant observations, and records containing inaccurate data. This study will demonstrate how to use simple pre-processing methods to improve the quality of input data.

Background. Although KDD has reported many successes in the business domain such as fraud detection and target marketing, the applications of KDD in the health care arena have been relatively few in comparison. This is primarily due to the problem of quality of data. Many health report data sets contain typographical errors, missing data, and duplicate records [1]. Goodwin *et al* also states that the issues obstructing progress in KDD for improved health outcomes include data quality problems, data redundancy, and data inconsistency [2]. In order for KDD to be used successfully, the quality of the input data must be improved through pre-processing while retaining as many cases as possible.

The purpose of this poster is to describe and illustrate the challenges and concerns in the data pre-processing step of an ongoing project, Developing Fall Predictive Model Utilizing KDD.

Methodology. This non-experimental study employs KDD to do secondary data analysis of the Minimum Data Set (MDS), which is a federally mandated comprehensive resident assessment instrument being used in all Medicare and Medicaid supported nursing homes in the United States, obtained from long term care (LTC) facilities in Kansas in 1996.

Setting and Sample: The targeted population is the elderly in LTC facilities during the year 1996. The file contains information on 405 LTC facilities, 37,897 LTC residents, 130,335 observations, and records of 58,816 falls.

Procedure. The pre-processing will proceed as follows:

First is the challenge of redundant observations. In this data set, one resident may have anywhere from 1 to 13 records. Redundant records present a source of error. Most data mining tools produce different results if some of the instances in data files are redundant, because repetition gives them more influence on the result [3]. To solve this problem, we select only the cases that include a first initial assessment to generate the predictive model. This assures that each unique individual will be represented by a single record.

The second challenge is dealing with missing data. In this data set, we found individual variables were missing data in anywhere from 10 to 100 percent of the records. If there are more than 50 percent of values for a variable, it becomes possible to use the *most common attribute value* method [2]. If not, the cases that have these missing data will be removed from the actual model building process.

Third is the challenge of inaccurate values. We had checked the data carefully for attribute values and found that some variables have an out-of-range value. These values will be replaced with the most common values from other records of that particular patient.

Results. Eliminating redundant observations reduced the number of cases by 45.8%. Because the MDS form had changed over time, seven variables contain no data. Therefore, these variables will be removed from this analysis. Applying the rule of missing data further reduced the remaining data by 57.1%. We are currently working on the inaccurate values and the final result will be presented in the poster.

Discussion. The quality of data is of great concern when applying KDD to health data provided in the public domain. Pre-processing is a necessary preliminary step before KDD can be used to successfully generate a model. Since there has been little investigation into exactly what one needs to do to pre-process the data, this study demonstrates how to apply three of the simplest pre-processing methods. Although these methods can eliminate redundant records, missing data, and inaccurate data, of primary concern is the reduction in the number of cases. This might affect the accuracy of the model. If the accuracy of the model is too low, then more sophisticated data pre-processing needs to be used to preserve a larger number of cases.

Acknowledgement*: The author wishes to acknowledge Dr. Patricia F. Brennan and the Center for Health Systems Research and Analysis for their assistance and support.*

Reference

[1] Hsu W. et al. (2000) Exploration data mining in diabetic patient database. *ACM*. 430-436.

[2] Goodwin et al (1997) Data mining issues for improved birth outcome. *Biomedical Science Instrumentation*. 34; 291-6.

[3] Witten I. & Frank E. (1997) *Data Mining*. San Francisco: Morgan Kaumann Publishers.